**PhD Synopsis**

**On**

# AN EFFICIENT APPROACH OF COMPRESSION FOR DATA WAREHOUSING


Submitted By

**Monika Sharma**

(Regn No: 08-ECM-3303)

Under the Supervision of

**Dr. Neera Batra**

**Associate Professor**

**Computer Science & Engineering Department Department**

M. M. Engineering College,Mullana

Ambala

**Department of Information Technology**

**M. M. Engineering College,**

**M. M. University, Mullana, Ambala, Haryana**


**April- 2014**

# AN EFFICIENT APPROACH OF COMPRESSION FOR DATA WAREHOUSING

## Ph.D Synopsis

### 1. INTRODUCTION

In the current scenario of changing business conditions organization's management needs to have access to more and better information. Most organizations are now days operating using information technology as the backbone of their operations but the fact is that despite having a large number of powerful desktop and notebook computers and a fast and reliable network, access to information that is already available within the organization is very difficult or otherwise not possible . All organizations whether large or small using Information Technology for the operations, produce large amount of data about their business including data about sales, customers, products, services and people. But in most cases, this data remains in the operational systems and can't be used by the organization. This phenomenon is called 'data in jail'. Experts say that only a small portion of this data that is entered, processed and stored is actually available to decision makers and management of the enterprise. The unavailability of this data can cause significant reduction in sales and profits of organizations and vice versa. In the 1990's, as large scale organizations began to need more timely data about their business, they found that traditional information systems technology was simply too slow and complex to provide relevant data efficiently and quickly. Completing reporting requests could take days or weeks using traditional reporting tools that were designed more or less to 'execute' the business rather than 'run' the business. As a cure for this problem, the concept of data warehouse started as a place where relevant data could be held for completing strategic reports for management. The key here is the word 'strategic' as most executives were less concerned with the day to day operations than they were with a more overall look at the model and business functions. In the latter half of the 20th century, there existed a large number and types of databases .Many large businesses found themselves with data scattered across multiple platforms and variations of technology, making it almost impossible for any one individual to use data from multiple sources. A key idea within data warehousing is to take data from multiple platforms/technologies and place them in a

common location that uses common querying tool. In this way, operational databases could be held on whatever system was most efficient for the operational business, while the reporting / strategic information could be held in a common location using a common language. Data Warehouses take this even a step farther by giving the data itself commonality by defining what each term means and keeping it standard. An example of this would be gender which can be referred to in many ways, but should be standardized on a data warehouse with one common way of referring to each sex. The purpose behind all these developments was to make decision support more readily available and without affecting day to day operations. One aspect of a data warehouse that should be stressed is that it is NOT a location for ALL of a businesses data, but rather a location for data that is 'interesting' and 'important'. Data that is interesting will assist decision makers in making strategic decisions relative to the organization's overall mission. Significant users of this technology include retail giants such as WalMart, credit card companies such as Visa and American Ex press, and major banks and transportation companies which include Bank of America, Royal Bank of Canada, Allied Irish Bank, United Airlines, Continental Airlines and many more. Planning the design and construction of these huge and complex information repositories have led to the development of data warehousing. Its major role remains crucial in understanding, planning, scoping and delivering knowledge capital back to the enterprise in a timely and cost effective fashion. Another important concept that has come out of the data warehousing concept is the recognition that there are two different types of information systems in all organizations namely operational systems and information systems. Operational systems are used to perform the day to day operations of the organization. They function as a backbone to any enterprise. For e.g. order entry, inventory control, payroll, accounting etc. are all operational systems. Because of their importance, the operational systems are always the first to be computerized in an enterprise. In fact, most of the organizations around couldn't operate without these operational systems. On the other hand, there are other functions within the organization which have to work with planning, forecasting and management. These functions are quite different from operational functions. For e.g., Resource planning, financial analysis and strategy planning etc. These functions require a lot of support from operational systems but these are actually different from operational systems. These are knowledge based systems called informational systems.

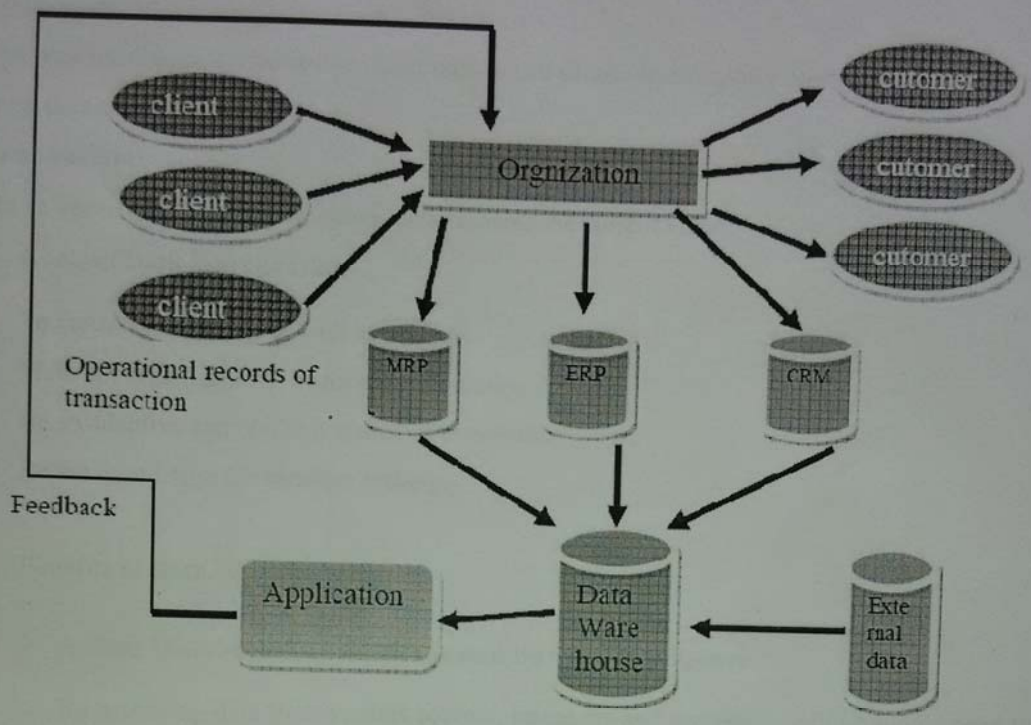


**Figure 1: Data Warehouse in an Organization**

Data Warehouse (The Data Warehouse) is a database of unique data structures that allows relatively quick and easy performance of complex queries over large amounts of data. A classical production information system is primarily adapted to input data. Since the production system requires the aforementioned property for such a system, it allows the company to be operational and run smoothly and that means mostly data entry.

The data warehouse is-

**Subject-oriented**

The data in the data warehouse is organized so that all the data elements relating to the same real-world event or object are linked together.

**Non-volatile**

Data in the data warehouse are never over-written or deleted — once committed, the data are static, read-only and retained for future reporting.

**Integrated**

The data warehouse contains data from most or all of an organization's operational systems and these data are made consistent.

**Time-variant**

For an operational system, the stored data contains the current value.

## 1.1 Goals of Data Warehousing

- To facilitate reporting as well as analysis.
- Maintain organizations' historical information.
- Be an adaptive and resilient source of information.
- Be the foundation for decision making.

## 1.2 Benefits of Data Warehouse

➢ **A Data Warehouse Delivers Enhanced Business Intelligence**

By providing data from various sources, managers and executives will no longer need to make business decisions based on limited data.

➢ **A Data Warehouse Saves Time**

Since business users can quickly access critical data from a number of sources—all in one place—they can rapidly make informed decisions on key initiatives. They won't waste precious time retrieving data from multiple sources.

➢ **A Data Warehouse Enhances Data Quality and Consistency**

A data warehouse implementation includes the conversion of data from numerous source systems into a common format.

➢ **A Data Warehouse Provides Historical Intelligence**

A data warehouse stores large amounts of historical data so you can analyze different time periods and trends in order to make future predictions.

Data compression is of interest in business data warehousing, both because of the cost savings it offers and because of the large volume of data manipulated in many business applications. The types of local redundancy present in business data files include runs of zeros in numeric fields, sequences of blanks in alphanumeric fields and fields which are present in some records and null in others. Run length encoding can be used to compress sequences of zeros or blanks. Null suppression may be accomplished through the use of presence bits. Another class of methods exploits cases in which only a limited set of attribute values exist. Dictionary substitution entails replacing alphanumeric representations of information such as bank account type, insurance policy type, sex, month etc. by a few bits necessary to represent the limited number of possible attribute values.

The problem of compressing digital data can be decoupled into two sub problems: modeling and entropy coding. Whatever the given data may represent in the real world, in digital form it exists as a sequence of symbols, such as bits. The modeling problem is to choose a suitable symbolic representation for the data and to predict for each symbol of the representation the probability that it takes each of the allowable values for that symbol. The entropy-coding problem is to code each symbol as compactly as possible, given this knowledge of probabilities. (In the realm of lossy compression, there is a third sub problem: evaluating the relative importance of various kinds of errors.)

## 2. LITERATURE SURVEY

Chen et al. [1] described that a data warehouse is an information provider that collects necessary data from individual source databases to support the analytical processing of decision-support functions. In the past, research on data warehouses primarily focused on relational data models. In this paper, the concept of object-oriented data warehousing is introduced and discussed. A new data model, called the compressed data model is proposed for storing the data in the object-oriented data warehouse. The data model will form new classes according to the definitions of views, such that the query performance and security can be improved. Three incremental maintenance algorithms, including instance insertion, deletion and update, are proposed to maintain the consistency between the data warehouse and the source databases.

Chen et al. [2] proposed a novel data model called the composite data model to store the data in the object oriented data warehouse. The data model forms new classes, consisting of the attributes listed in the definitions of views and copies necessary class structures from the data

source. The query performance of the data warehouse can thus be improved. The corresponding view creation and deletion algorithms are also proposed.

A novel data model for constructing an object oriented data warehouse has been proposed by Shieh et al. [3] which importantly preserves the original inherited hierarchies of the data sources. It also makes the data warehouse concise and simple thus improves the efficiency in data retrieval.

Abhijit et al. [4] has proposed the concept of using various database operators that permit to enrich technique of query optimization existing in the object oriented database based on cost, cardinality and number of bytes. Focus is on query optimization using relational operators, logical operators and special operators. This paper shows that the improvement in the quality of plans is significant only with decrease in cost, cardinality and number of bytes after compression.

Balakrishna et al. [5] described that increasing test costs have been one of the disadvantageous consequences of technology scaling especially in deep sub-micron designs. The amount of test data required to achieve good test quality has increased tremendously due to the increasing complexity of devices as well as the need to test for newer defect mechanisms that are becoming predominant in smaller device geometries. This has led to the development and deployment of new design-for-test (DFT) technologies to mitigate the problem. Test data compression has been at the forefront of solutions to reduce test costs through reduction in tester storage and test application time. Most test data compression techniques have concentrated on scan test vectors since the bulk of the increase in test data is due to scan vectors, including both stuck-at and delay tests.

Data compression is a necessary technique required in various scenarios from data communication to data storage these days. Text is an important form of data used ubiquitously in different communications and in computer world. Fahad et al. [6] presents a novel data compression technique that uses an evolutionary programming approach for the compression process. Text is used as the experimental data in this research. By using evolution, the best compression method(s) are chosen in order to achieve maximum compression accuracy. For different experiments, the compression extent is measured and also the results are compared with the compression methods, individually. The results reveal the commendable performance of the system and the effect of evolution on the overall compression.

Jonathan et al. [7] proposed a simple lossless visual image compression scheme. In this scheme, the two dimensional visual image data is converted to a one dimensional data using our proposed pixel scanning method. The difference between consecutive pixel values in the resulting one dimensional image data is taken and the residues are encoded loss lessly using an entropy encoder. The working principle of this approach is presented together with the image compression algorithm used. A software algorithm is also developed and implemented to compress some standard test images using Huffman style coding techniques in a MATLAB platform.

S. H. Kim et al. [8] suggested that in huge video databases, an effective video indexing method is required. While manual indexing is the most effective approach to this goal, it is slow and expensive. Thus automatic indexing is desirable, and previously various indexing tools for video databases have been developed. For efficient video indexing and retrieval, the similarity measure is an important factor. This paper presents new similarity measures between frames and proposes a new algorithm to detect scene changes using a cross entropy defined between two histograms. Experimental results show that the proposed algorithm is fast and effective compared with several conventional algorithms to detect abrupt scene changes and gradual transitions including fade in/out and flash light scenes.

Liu et al. [9] presented an efficient compression-oriented segmentation algorithm for computer-generated document images. In this algorithm, a document image is represented in a block-based multi-scale pyramid. Then, image blocks will be characterized based on their entropy values of the intensity histogram and the entropy distribution is assumed to be Gaussian priors in this work. Two methods off-line and online training has been used to estimate model parameters. The multi-scale Bayesian estimation has been used to refine the classification results and generate the final segmentation result where image blocks are classified into four classes i.e. background, text, graphic and picture. It is expected that the proposed entropy-based segmentation will be suitable for compound document compression and two training approaches apply to different applications.

Muthu kumar et al. [10] invented a new technique to analyze the database compression performance for real time database system. This proposed technique not only reduces space requirements on disk and input/output performance, it also reduces the utilization of memory,

thus reducing the number of buffer faults resulting in in input/output. This algorithm enables granular, enhances the compression and decompression performance.

In another scenario, the lossless method of image compression and decompression using the simple coding technique called Huffman coding has been proposed [11]. This technique is simple in implementation and utilizes less memory. A software algorithm has been developed and implemented to compress and decompress the given image using Huffman coding technique in a MAT LAB platform.

Anubhuti et al. [12] presented an efficient and minimum hardware implementation technique. Voice data compression and decompression is about a process which reduces the data rate or file size of digital audio signals. This Process reduces the dynamic range of audio signals.

Prateek et al. [13] distributed the medical images to different hospitals and among the staff of the same medical center within the short span of time and efficiently a lot of hospitals handle their medical image data with computers. To overcome the problems concerned with image compression, filmless imaging and digital compression techniques are used..

Raja et al. [14] proposed two level text compression and decompression techniques for lossless data compression. The features of both LZW (Lempel-Ziv-Welch) and Huffman algorithms are combined to improve the compression ratio. The main advantage of this combined algorithm is that the percentage of data increases more than five percent compared to the existing text compression technique.

Shrusti et al. [15] provides lossless data compression methodologies and compares their performance. Huffman and arithmetic coding are compared according to their performance. Data Compression is a process that reduces the data size, removing the excessive information. Shorter data size is suitable because it is simply reducing the cost. The aim of data compression is to reduce the redundancy in stored or communicated data, thus increasing effective data density. There are number of different data compression methodologies which are used to compress different data formats like text, video, audio, image files. There are two forms of data compression "Lossy" and "Lossless" in lossless data compression the integrity of data is preserved.

S.J.O et al. [16] examined the performance gains to be made by compression outside the index. A novel compression algorithm is reported, which enables the processing of queries without decompressing data needed to perform join operation in database built on a triple store. It is

found that for some applications, gains in performance of over 50% are achievable, and in OLTP like solutions, there are also gains to be made.

Pabbisetti et al. [17] used three methods for both data compression and decompression process, (i) discrete Hartley type transform (ii) fast Fourier transform (FFT) (iii) discrete cosine transforms(DCT). Algorithms are developed and tested using the three methods on different images. A comparison with respect to mean square error with the original image also presented. The main advantage of discrete Hartley type transform is, it is real transform. Discrete cosine transform also has similar performance as that of discrete Hartley type transform. An algorithm for compression using FFT method is also presented.

A novel method for the compression and decompression of digital color images has been proposed by Santanu et al. [18]. A novel algorithm has been proposed by which a color image can be decompressed which is compressed using compression algorithm. The results show that the coding performance and the estimated time required for compressing an image and decompressing an image is significantly increased.

Naresh et al. [19] implemented compression in traditional database system to improve the performance significantly, The proposed method reduces the size of the data and improves input/output performance by reducing seek times, reducing transfer times and increasing buffer hit rate. In a column oriented database, compression schemes that encode multiple values at once are natural. In a row oriented database such schemes don not work as well because an attribute is stored as a part of entire tuple, so combining the same attribute from different tuples together into one value would require some way to "mix" tuples.

Dalvir kaur et al. [20] presented a columnar oriented database enhancing approach to service the needs of business intelligence. One of the major advantage of column based database approach is that it is easy to understand its effect on compression and possible to apply compression algorithm for each column. They used the compression on columnar oriented database using hybrid approach. This consists of two lossless algorithm Huffman coding and Lempel-Ziv-welch coding in MATLAB software.

## 3. PROBLEM FORMULATION

It is evident from the literature survey that following gaps exist in the present study:-

- There is need to plan to improve the efficiency of databases.
- Introducing the compression and decompression scheme.

- Increasing the compression ratio and decreasing the query processing time.

## 4. OBJECTIVES

The objectives of the present study are as follows:-

- To develop a lossless compression and decompression algorithm at attribute level in order to achieve a reasonable compression ratio.
- To develop efficient compression algorithm in data warehousing to improve the efficiency of the data warehousing packages so that less CPU time and less Memory is consumed.
- Comparison of time taken and compression efficiency for different sizes of databases.
- To Implement compressor and expander using compression algorithm and test its effectiveness on different sized databases to save disk space and less time in data warehouse.
- Increasing the compression ratio and decreasing the query processing time through compression algorithm.

## 5. METHODOLOGY

Following steps will be taken for the thesis work

- Methodology: High Level programming language either Java will be used for implementation.
- Database: Sample Database containing different types of text or any other type of data will be created and extracted for storage in data warehouse.

## References


[1] W.C. Chen, T.P. Hong and W.Y. Lin, "Using The Compressed Data Model In Object-Oriented Data Warehousing", In the Proceedings of IEEE on Systems, Man, and Cybernetics, vol.5, pp.768-772, 1999.

[2] W.C. Chen, T.P. Hong and W.Y. Lin, "A Composite Data Model In Object-Oriented Data Warehouses", In the Proceedings of the 31st International Conference on Technology of Object-Oriented Language and Systems, pp. 400-405, 1999.

[3] J.C. Shieh and H.W. Lin, "The Novel Model of Object-Oriented Data Warehouses", Workshop on Databases and Software Engineering, Taiwan, 2006.

[4] A.Banubkode and H.Acharya, "Query Optimization on Compressed and Decompressed Object Oriented Database Using Operators", In the Proceedings of International Journal of Computer Science and Engineering, Vol.2, No. 9, pp.2883-2889, 2010.

[5] K. J. Balakrishna and N. J. Princeton, "Emerging Techniques For Test Data Compression", In the Proceedings of IEEE on 14th Asian Test Symposium at Calcutta, pp.462, 20-21, Dec, 2005.

[6] F. Ullah and K. M. Yahya, "A New Data Compression Technique Using an Evolutionary Programming Approach", Emerging Trends and Applications in Information Communication Technologies Communications in Computer and Information Science, Vol. 281, pp.524-531, 2012.

[7] J. G. Kolo, K. P. Seng, L.M. Ang and S. R.S. Prabaharan, "Data Compression Algorithms for Visual Information", In the Proceedings of International Conference of Informatics Engineering And Information Science, Vol. 253, pp. 484-497, 2011.

[8] S. H. Kim and R.H. Park, "A Novel Approach To Scene Change Detection Using A Cross Entropy", In The Proceedings Of International Conference On Image Processing, vol.3, pp. 937-940, 2000.

[9] L. Liu, Y. Dong, X. Song and G. Fan, "An Entropy Based Segmentation Algorithm For Computer-Generated Document Images", In The Proceedings of International Conference on Image Processing, Vol. 1, pp. 541-544, 2003.

[10] M. Muthukumar, T. Ravichandran, "Analysing Compression Performance For Real Time Database System", In The Proceedings of Journal of Theoretical and Applied in Information Technology, Vol. 1, No.2, pp.17-22, 2012.

[11] Jagdish H. Pujar, Lohit M. Kadlaskar, "A New Lossless Method Image Compression and Decompression using Huffman Coding Technique", Journal of Theoretical and Applied Information Technology, Vol.15, No.1, pp.18-23, 2005.

[12] A. Khare, M. Saxena, and B., Patil, "Voice data compression and decompression", In The International Journal of Engineering and Advanced Technology, Vol.1, No.1, pp.6-10, , 2011.

[13] P. Verma, P. Verma, A. Sahu, S. Sahu and N. Sahu, " Comparison between Different Compression And Decompression Technique On MRI Scan Images", International Journal

of Advanced Research in Computer Engineering And Technology ,Vol.1, No.7, pp.109-113, 2012.

[14] Raja, P. and Saraswathi, D.,"An Effective Two Stage Text Compression And Decompression Technique For Data Communication", International Journal of Electronics and Communication Engineering, Vol.4, No.2, pp. 233-241, 2011.

[15] Shrusti Porwal, Yashi Chaudhary, Jitendra Joshi and Manish Jain, "Data Compression Methodologies For Lossless Data And Comparison Between Algorithms", International Journal of Engineering Science and Innovative Technology, Vol.2, No.2 , pp.142-147, 2013.

[16] S.J.O. Connell and N. Winterbottom, "Perform Joins without Decompression in a Compression Database System", In The SIGMOD Record, Vol.32, No.1, pp.6-11, 2003.

[17] P. Sathyanarayana, and Hamid R. Saeedipour, "Digital Image Compression and Decompression Using Three Different Transforms and Comparison of Their Performance", In the Proceedings of International Conference on Man Machine System at Malaysia, pp.1-5, 15-16, Sept, 2006.

[18] S. Santra, and P. P. Acharjya, "A novel approach of digital image compression technique using bit plan size reduction", International Journal of Current Engineering and Technology, Vol. 3, No.4, pp.1491-1494, 2013.

[19] Naresh Kumar and K. K. Bansal, "Different compression techniques and their execution in database systems to improve performance", International Journal of Advanced Research In Computer Science and Software Engineering, Vol. 2, No. 6, pp.293-296, 2012.

[20] Dalvir Kaur and Kamaljeet Kaur, "Data Compression on Columnar–Database Using Hybrid Approach (Huffman And Lempel- Ziv- Welch Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, No. 5, pp.889-894, 2013.